UNCOVERING THE GENERATIVE LANDSCAPE OF DIFFUSION MODELS: A SPECTRAL-GUIDED QUALITY-DIVERSITY DIFFUSION FRAMEWORK

Anonymous authors
Paper under double-blind review

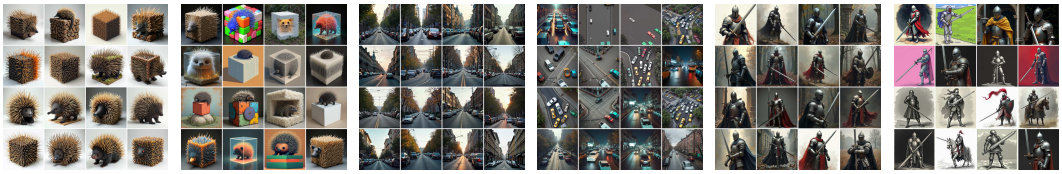


Figure 1: Flux.1 [dev] (Labs, 2024) outputs under standard sampling versus our Spectral-Guided Quality-Diversity Diffusion. (Left) "a cube made of porcupine", (Middle) "a street with several cars on it", and (Right) "a knight holding a long sword". Random sampling converges to structurally similar modes. Our method exhibits more visual diversity while maintaining prompt alignment.

ABSTRACT

Denoising Diffusion Probabilistic Models have emerged as powerful generative priors across domains ranging from molecular design to high-resolution image synthesis. However, standard diffusion sampling is inherently biased toward high-likelihood regions, often limiting the diversity of outputs. In many domains, a single, high-likelihood sample is insufficient. Instead, diverse portfolios of high-performing solutions are required, motivating methods that uncover a model's full expressive range. Here we show that integrating Quality-Diversity (QD) optimisation with diffusion models enables the discovery of diverse, high-performing outputs. We first propose *Quality-Diversity Diffusion*, a framework that adapts QD to diffusion models by directly optimising the model's latent noise and conditioning embeddings and leveraging embedding models to quantify diversity. Second, we introduce a novel h*-space Spectral Mutation* operator, which guides search along directions of maximal semantic sensitivity by analysing principal singular vectors of the model's input-to-activation Jacobian. We evaluate our combined approach, *Spectral-Guided Quality-Diversity Diffusion*, on text-to-image generation. Against 12 baselines, our method consistently achieves the lowest FID score, demonstrating an excellent balance of image quality and diversity.

1 INTRODUCTION

Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) have emerged as powerful generative priors, achieving state-of-the-art results across domains from protein design (Watson et al., 2023) to audio synthesis (Kong et al., 2020). By learning to reverse a gradual noise-adding process, diffusion models approximate the underlying data distribution of a given design domain.

However, standard diffusion sampling procedures are inherently biased towards high-likelihood regions of the learned distribution, often suppressing the diversity inherent in the model's latent space (Ho and Salimans, 2022; Sadat et al., 2024; Corso et al., 2023). In many design tasks, a single high-likelihood solution is insufficient: in biological or materials design, candidates that perform well *in silico* may fail under real-world conditions due to modelling errors or the sim-to-real gap. Instead, identifying a *diverse portfolio* of structurally distinct candidates increases the chance that at

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

least one succeeds in downstream experimental validation (Dona et al.; Janmohamed et al., 2024; Wolinska et al., 2024). Alternatively, in text-to-image generation, diverse outputs support creative exploration by exposing users to a range of possible visual interpretations of a prompt (Resnick et al., 2005). These examples highlight the need for methods that uncover the full expressive range of a generative prior, rather than just its dominant modes.

In principle, diffusion models generate diverse outputs by resampling the initial noise; in practice, large-scale models remain biased towards the modes of the training distribution, frequently producing artefacts with near-identical global structure across seeds (see Figure 1). This calls for a principled mechanism for *exploration* rather than repeated exploitation of high-probability regions – a role well suited to Quality-Diversity (QD) optimisation (Cully and Demiris, 2018; Pugh et al., 2016). Unlike traditional optimisation, which converges to a single global optimum, QD algorithms illuminate the search space returning an archive of high-performing solutions that are behaviourally diverse.

Applying QD optimisation to diffusion-parameterised tasks presents two key challenges. First, diversity must capture meaningfully semantic or functional differences between artefacts (i.e., images, molecules, audio). In many complex domains, such notions of diversity are difficult to specify manually and are highly task-dependent. We address this by leveraging pretrained foundation models, whose embedding spaces act as proxies for distinctiveness between artefacts. However, their high dimensionality makes them difficult to integrate into existing QD approaches. Second, the search space induced by a diffusion model is itself high-dimensional, requiring search strategies that can efficiently navigate it while producing meaningful variation in the generated artefacts.

In this work, we introduce *Spectral-Guided Quality-Diversity Diffusion*, a new approach for QD algorithms designed to uncover the hidden diversity of diffusion models. Our contributions are:

1. **QD-Diffusion:** We present a Quality-Diversity framework that directly optimises the continuous latent noise and conditional embedding spaces of diffusion models. Our approach measures diversity via high-dimensional embeddings from pretrained foundation models. By doing so, our framework avoids hand-defined, low-dimensional features required by previous QD approaches, enabling application to domains where explicit axes of diversity are difficult to specify.

2. **The h-space Spectral Mutation Operator:** We introduce a novel mutation operator tailored to the geometry of diffusion models. Recognising that traditional evolutionary operators in high-dimensional latent spaces are inefficient, our operator computes the principal singular vectors of the model's input-to-activation Jacobian. Perturbing along these directions of maximum semantic sensitivity enables efficient exploration of the latent manifold.

We validate our approach on the challenging domain of text-to-image generation using a representative subset of the Parti Prompts benchmark (Yu et al., 2022). We test our framework against 5 distinct baseline algorithms and benchmark our mutation operator against 7 alternative operators, demonstrating that our method achieves the lowest FID score across all 12 baselines. Our code is open-source, available at *Code hidden for anonymity, will be released upon acceptance.*

2 Background

2.1 Diffusion Models

Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) are a class of generative models that learn to approximate a data distribution $q(\mathbf{x}_0)$ by reversing a gradual noise-adding process. The framework is defined by two Markov chains: a fixed forward process that gradually perturbs data with Gaussian noise until it becomes indistinguishable from isotropic noise, and a parametrised reverse process that learns to denoise it.



Figure 2: Visualisation of the reverse diffusion process.

The Reverse Generative Process Once the model, parametrised by θ, is trained, the core of the generative capability lies in the reverse Markov chain. To generate an image, we sample pure noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively denoise it over T timesteps, as visualised in Figure 2.

The joint distribution of this reverse process, parametrised by θ, is defined as:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\,, \tag{1}$$

where the transition kernel $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ represents a single denoising step. Assuming the noise added at each step of the forward process is sufficiently small, this reverse transition can be approximated by a Gaussian distribution:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \tag{2}$$

At each step, the model estimates the mean $\boldsymbol{\mu}_\theta$ and variance $\boldsymbol{\Sigma}_\theta$ required to transition from the noisy state $\mathbf{x}_t$ to the less noisy state $\mathbf{x}_{t-1}$. By recursively applying this transition starting from T, the model gradually resolves structure, resulting in a sample $\mathbf{x}_0$ from the learned data distribution. Conditional generation is achieved by injecting context embeddings c into the denoising network via cross-attention or concatenation, guiding outputs toward the desired attributes.

2.2 Quality-Diversity Algorithms

Quality-Diversity (QD) algorithms aim to illuminate a search space by discovering a collection of solutions that are both high-performing and behaviourally diverse (Cully and Demiris, 2018). In the QD paradigm, each candidate solution x within the search space $\mathcal{X}$ is evaluated by two functions: a fitness function $f : \mathcal{X} \to \mathbb{R}$, which measures task performance, and a feature function $\phi : \mathcal{X} \to \Phi(\mathcal{X}) \subset \mathbb{R}^d$, which maps the solution to a feature space $\Phi(\mathcal{X})$ where diversity is sought. The objective is to find a set of solutions $\mathcal{A} = \{x^{(1)}, \ldots, x^{(N)}\}$ that maximises $f(x)$ for every unique region or feature in $\Phi(\mathcal{X})$. For example, in materials design, the fitness f might measure stability while the feature ϕ captures conductivity; QD then seeks a set of stable materials spanning diverse conductivity values (Wolinska et al., 2024; Janmohamed et al., 2024).

The standard QD framework maintains an archive of solutions organised by their features. At each iteration, candidate solutions are selected from the archive, subjected to variation operators (mutation or crossover), evaluated, and potentially inserted back into the archive based on an elitist replacement rule (Cully and Demiris, 2018). A canonical instantiation is MAP-Elites (Mouret and Clune, 2015), which discretises the feature space $\Phi(\mathcal{X})$ into a multi-dimensional grid. A new candidate x' is added to a cell if that cell is empty, or if $f(x')$ exceeds the fitness of the existing occupant $f(x)$. This mechanism ensures the archive monotonically improves in quality while maintaining diversity across the grid. We provide the pseudocode for MAP-Elites in Appendix A.1.

While effective in low-dimensional settings, the grid-based discretisation of MAP-Elites becomes inefficient as the dimensionality of $\Phi(\mathcal{X})$ increases. Dominated Novelty Search (DNS) (Bahlous-Boldi et al., 2025) overcomes this by maintaining an unstructured archive. At each generation, both offspring and archived solutions are ranked using a scoring metric that balances fitness and novelty. Specifically, a solution's score is calculated as the average distance in feature space to its k-nearest neighbours that possess higher fitness. This logic ensures that solutions are preserved if they are either distinct (isolated in feature space) or high-performing (few neighbours have better fitness). Because DNS relies on distance metrics rather than grid discretisation, it scales effectively to high-dimensional feature spaces. The pseudocode for DNS is provided in Appendix A.2.

3 Methodology

3.1 Framework

Overview In this section, we detail the *Quality-Diversity Diffusion framework* which integrates diffusion models with evolutionary search to discover diverse, high-performing designs. We assume the existence of a design task defined by:

1. A pre-trained diffusion model $p_\theta(x_{0:T})$ capable of generating data samples.
2. A fitness function $f : \mathcal{X} \to \mathbb{R}$ that evaluates the quality or task performance of a generated sample.

3. An embedding model $\phi : \mathcal{X} \to \mathbb{R}^d$ that maps samples to a feature space representing semantic or behavioural characteristics.

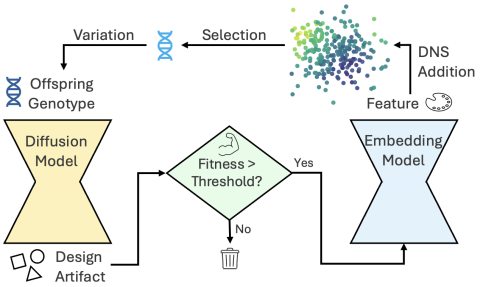


Figure 3: Quality-Diversity Diffusion Framework.

Our framework follows a standard Quality-Diversity loop (see Figure 3). Genotypes (e.g., latent noise x_T and conditioning embeddings c) are selected and varied to produce offspring. In this work, we propose the h-space Spectral Mutation Operator although any variation operator can be used. The diffusion model maps these to new artifacts, which are evaluated for fitness, embedded in a high-dimensional feature space and added to the unstructured archive using Dominated Novelty Search with fitness thresholding.

Genotypes and Mutations We define the genotype as the tuple (x_T, c), where $x_T \sim \mathcal{N}(0, I)$ is the initial Gaussian noise vector and c is the conditioning embedding vector. Restricting our search space to these variables ensures the framework is domain-agnostic and applicable to any diffusion-based task (e.g., images, proteins, audio) without requiring text-specific linguistic priors.

DNS With Fitness Thresholding Standard QD algorithms such as MAP-Elites often rely on low-dimensional, hand-crafted features (Chatzilygeroudis et al., 2021; Cully and Demiris, 2018; Mouret and Clune, 2015) which are often unavailable in complex domains. In this work, we rely on the assumption that an increasing number of design domains have access to rich representations via pre-trained embedding models (e.g., CLIP for images (Radford et al., 2021), ESM for proteins (Lin et al., 2023)), which act as powerful proxies for semantic distinctiveness. However, the high dimensionality of these vectors (often hundreds of dimensions) renders grid-based algorithms infeasible.

To address this, we employ Dominated Novelty Search (DNS) as the underlying archive mechanism. As explained in Section 2.2 DNS maintains an unstructured archive, calculating a score for each solution based on the average distance to its k-nearest fitter neighbours in feature space. While DNS has been shown to scale well to high-dimensions, we identified a critical limitation when applying it to generative tasks. In particular, since the DNS scoring mechanism balances the quality and novelty of a solution, we found that extremely novel solutions were valued highly regardless of their validity. In generative design, mutations in the latent space can easily traverse boundaries into invalid regions of the manifold (e.g., generating a live "turkey" instead of a "roast turkey"); standard DNS erroneously preserves these failures due to their high novelty scores. Hence, degenerate solutions are frequently retained and then propagated through the archive solely because they are novel. To mitigate this, we introduce a fitness-thresholding mechanism. A candidate solution x' is only considered for addition to the archive if its fitness exceeds a minimum threshold, i.e., $f(x') > f_{\text{thresh}}$. This restricts search to valid regions of the generative manifold. An ablation study on this mechanism is provided in Appendix Section D.1.

3.2 h-SPACE SPECTRAL MUTATION OPERATOR

Given the framework outlined in Section 3.1, the primary challenge shifts to efficient exploration within the vast latent space of diffusion models. To address this, we introduce the **h-space Spectral Mutation Operator**.

Spectral Sensitivity Analysis Our operator aims to traverse the latent space along meaningful axes, which requires extracting the latent structure of the input space $\mathcal{X}_t$. Directly analysing $\mathcal{X}_t$ is difficult, however, since diffusion training does not enforce a globally organised semantic structure, and the geometry of $\mathcal{X}_t$ varies with timestep t (Park et al., 2023).

We instead turn to the model's internal bottleneck representations, or *hidden space* $\mathcal{H}$ (h-space), which are known to preserve local linear structure corresponding to semantic features (Kwon et al., 2022). Let $g : \mathcal{X}_t \to \mathcal{H}$ denote the mapping from the latent at timestep t to this hidden space. Since g is non-linear, we characterise its local geometry via *spectral analysis* of its Jacobian, using

differential geometry to approximate its behaviour between tangent spaces $\mathcal{T}_{\mathcal{X}}$ and $\mathcal{T}_{\mathcal{H}}$. Following Park et al. (2023), the Jacobian $J_g(\mathbf{x}) = \nabla_{\mathbf{x}} g(\mathbf{x})$ defines a linear map $\mathcal{T}_{\mathcal{X}} \to \mathcal{T}_{\mathcal{H}}$, letting us identify the perturbation $\mathbf{v} \in \mathcal{T}_{\mathcal{X}}$ that induces maximal variance in $\mathcal{T}_{\mathcal{H}}$:

$$\mathbf{v}^* = \arg\max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \|J_g(\mathbf{x})\mathbf{v}\|^2 \tag{3}$$

$\mathbf{v}^*$ is the first right singular vector of $J_g(\mathbf{x})$ (i.e., the first column of V in $J_g(\mathbf{x}) = U\Sigma V^T$, satisfying $J_g(\mathbf{x})\mathbf{v}^* = \sigma_1 \mathbf{u}_1$): moving along it produces the largest possible change in the model's internal features for a fixed step size.

Based on this analysis, we define our mutation operator as a perturbation of the latent state $\mathbf{x}_t$ along this principal semantic direction:

$$\tilde{\mathbf{x}}_t = \mathbf{x}_t + \alpha \mathbf{v}^* \tag{4}$$

where $\alpha \sim \text{Unif}[\alpha_{\min}\alpha_{\max}]$, is the step size. By aligning perturbations with these principal axes, we force the search along the most semantically active regions of the latent manifold, avoiding the inefficiencies of unguided evolutionary variation operators. We note that this approach is architecture-agnostic: $g(\mathbf{x}; t, c)$ is a generic input-to-feature mapping, corresponding to an intermediate convolutional block in U-Net models and a transformer layer in DiT models, so the analysis applies regardless of backbone.

Furthermore, we extend the analysis of Park et al. (2023) beyond the latent noise $\mathbf{x}_t$, by applying the same spectral analysis to the conditioning vectors c, identifying directions of maximal sensitivity in the conditional embedding space. This allows our framework to treat conditioning vectors as mutable genotypes alongside $\mathbf{x}_T$, enabling joint optimisation of the seed and the conditional embeddings.

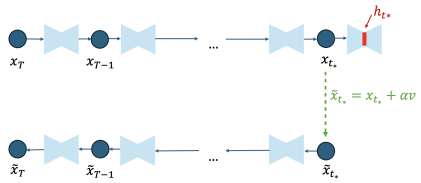


Figure 4: The h-space Spectral Mutation Operator. Initial noise $\mathbf{x}_T$ is propagated to a timestep t_* where singular vectors $\mathbf{v}$ of the Jacobian mapping the latent $\mathbf{x}_t$ to the model's internal activations h are computed. The latent is perturbed along this direction ($\tilde{\mathbf{x}}_t = \mathbf{x}_t + \alpha\mathbf{v}$) and the result $\tilde{\mathbf{x}}_t$ is deterministically inverted to the timestep to form the offspring genotype $\tilde{\mathbf{x}}_T$.

Deterministic Inversion Evolutionary search on generative models is typically performed directly on the initial noise vector space $\mathcal{X}_T$, i.e., at $t = T$ (Sobania et al., 2025). However, literature in diffusion editing suggests that manipulating latent variables at this initial stage ($t = T$) rarely yields meaningful semantic shifts. Instead, semantic editing is most effective at intermediate timesteps (typically $t_* \in [0.5T, 0.7T]$), where internal representations are most distinct (Park et al., 2023). However, we cannot directly store intermediate latents in the archive, since genotypes from different diffusion timesteps would be incompatible for subsequent parent selection. We resolve this by mutating at an intermediate timestep t_* and using deterministic noise schedulers to invert the modified latent back to the initial noise space $\mathbf{x}_T$, keeping the genotype consistent for archiving. The overall process, visualised in Figure 4, proceeds as follows:

1. **Forward Simulation:** The parent genotype $\mathbf{x}_T$ is propagated to an intermediate timestep t_* (randomly sampled or a hyperparameter) using a deterministic discrete scheduler such as DDIM or Euler Discrete (Ho et al., 2020; Karras et al., 2022). In these deterministic settings, the variance term $\mathbf{\Sigma}_\theta$ (which typically controls stochastic noise injection) is effectively zero, ensuring that the trajectory is bijective. The intermediate latent is thus obtained by iteratively applying the scheduler's deterministic update rule, using only the predicted mean $\boldsymbol{\mu}_\theta$:

$$\mathbf{x}_{s-1} = \boldsymbol{\mu}_\theta(\mathbf{x}_s, s) \quad \text{for } s = T, \ldots, t_* + 1 \tag{5}$$

2. **Perturbation:** Since exact computation of $J_g(\mathbf{x})$ is intractable, we follow Park et al. (2023) and employ a *Power Iteration* method leveraging finite differences and automatic differentiation to

approximate the top-k singular vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ of the mapping g's Jacobian. We compute these sensitivity directions for a single parent and create a batch of b new solutions by sampling different scalings α and vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$:

$$\tilde{\mathbf{x}}_{t_*} = \mathbf{x}_{t_*} + \alpha \mathbf{v} \tag{6}$$

where α is a scalar step size sampled uniformly from $[\alpha_{\min}, \alpha_{\max}]$ and $\mathbf{v}$ is randomly chosen from $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, reusing the expensive gradient computation across the whole offspring batch.

3. **Deterministic Inversion:** Since deterministic schedulers define a bijective mapping between noise levels, we can compute the inversion by iteratively applying the inverse update rule:

$$\tilde{\mathbf{x}}_{s+1} = \boldsymbol{\mu}_\theta^{-1}(\tilde{\mathbf{x}}_s, s) \quad \text{for } s = t_*, \ldots, T-1 \tag{7}$$

translating semantic changes at t_* back into the noise space $\mathbf{x}_T$. This step is unnecessary for distilled, one-step models, where perturbations apply directly to the input noise.

4 Experimental Setup

4.1 Task

Domain We assess the efficacy of our method in the domain of text-to-image generation. We conduct our primary evaluation on SDXL-Turbo (Sauer et al., 2024), selected for its efficiency. To ensure our findings hold across diverse semantic contexts, we utilise the Parti Prompts benchmark (Yu et al., 2022). For SDXL-Turbo, we sample three prompts from each of the provided categories (e.g., abstract, complex), resulting in a diverse set of 36 evaluation tasks. To demonstrate the generality of our framework, we conduct additional experiments using the FLUX.1 [dev] model (Labs, 2024) which utilises a Diffusion Transformer (DiT) architecture and relies on a multi-step generation process (typically 30-50 steps), in contrast to the U-Net-based, single-step generation of SDXL-Turbo. Due to the substantial computational cost associated with the Flux model, we restricted this evaluation to a subset of 12 prompts. Further implementation details are provided in Appendix Section C.

Fitness Definition The fitness is the fidelity of the images to the prompt, quantified via the CLIP score (Radford et al., 2021), a standard metric for image-text alignment.

Feature Definition We use DreamSim (Fu et al., 2023) for our features. The model generates 768-dimensional embeddings trained to approximate human similarity judgements, thereby serving as a proxy for semantic distinctiveness.

Genotype As noted in Section 3.2, the spectral variation operator is general and can operate on any genotype. For each model, we define the genotype to encompass all available input components, splitting the batch size equally among them. We apply the same genotype definitions to all applicable baselines. An ablation study isolating each genotype component is provided in Section D.2.

4.2 Baselines

An overview of each baseline is provided below. A comprehensive related works and implementation details of baselines are provided in Appendices Section B and Section C.3 respectively.

Framework Baselines We compare the QD-Diffusion framework design to the following baselines:

- **CVT-ME:** A variant of MAP-Elites designed to scale to high-dimensional spaces by using Centroidal Voronoi Tessellations to discretise the feature space into a fixed number of cells (Vassiliades et al., 2018).
- **CMA-ES:** A robust evolutionary optimisation strategy designed for high-dimensional spaces (Hansen and Ostermeier, 1996). Unlike QD algorithms, CMA-ES does not use a feature space and aims solely to maximise fitness.
- **CMA-MAE:** A state-of-the-art Quality-Diversity algorithm that combines CMA-ES with the MAP-Elites archiving mechanism (Fontaine and Nikolaidis, 2023). While traditionally reliant on hand-defined feature spaces, we adapt this baseline to use a CVT-ME grid over DreamSim embeddings to ensure a fair comparison within our automated, high-dimensional feature setting.
- **Adam:** Gradient-based optimisation of the latent variable over the fitness function (Kingma, 2014).
- **QDHF:** Quality-Diversity via Human Feedback (Ding et al., 2023). This method uses CLIP embeddings as features and a learned projection layer to map high-dimensional embeddings to a low-dimensional latent space suitable for standard MAP-Elites. The learned layer is trained using a contrastive loss for similarity, with DreamSim serving as a proxy for human judgements.

The CVT-ME, CMA-MAE and QDHF baselines are selected to determine whether our framework is the optimal QD algorithm for high-dimensional feature embeddings. Crucially, CVT-ME and QDHF employ our proposed spectral mutation operator, ensuring that comparisons isolate the efficacy of the archive mechanism rather than the variation strategy. CMA-ES and Adam represent pure fitness-maximising algorithms, while serving to test whether an explicit diversity mechanism is necessary or if powerful optimisers can implicitly discover diversity. Adam and QDHF are not evaluated on Flux due to computational constraints (see Appendix Section C.3).

Variation Operator Baselines To isolate the efficacy of the h-space Spectral Mutation operator from the framework, we benchmark it against the following variation operators:

- **Random seeds:** A baseline wherein each iteration samples a batch of solutions via new random seeds, representing the standard usage pattern of diffusion models.
- **Diverse-Diffusion:** A recent approach designed to enhance image variety by performing a search for latent vectors with maximum pairwise distance (Zameshina et al., 2023).
- **Particle Guidance:** generates a batch of samples jointly rather than i.i.d., adding a guidance term that repels particles from one another to enforce within-batch diversity (Corso et al., 2023).
- **ImageBreeder:** An evolutionary operator originally designed for image generation that performs latent-space crossover using a variety of strategies, including weighted blending and slicing along channels or spatial dimensions (Sobania et al., 2025). This is the only operator that is only applied to the latent noise vectors, as the operator was developed specifically for latent-space crossover.
- **Gaussian Noise:** Adds isotropic perturbations to the genotype, defined as $\mathbf{x}' = \mathbf{x} + \sigma\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and σ is the mutation strength hyperparameter.
- **Isoline-DD:** Generates offspring by perturbing a parent $\mathbf{x}_1$ along the direction relative to a second parent $\mathbf{x}_2$, defined as $\mathbf{x}' = \sigma_1\mathbf{x}_1 + \sigma_2\mathcal{N}(0,1)(\mathbf{x}_1 - \mathbf{x}_2)$ where σ_1 and σ_2 are hyperparameters (Vassiliades and Mouret, 2018).
- **Intermediate-Layer Gaussian noise:** Instead of performing SVD-based updates at the intermediate timestep, we apply Gaussian noise to the intermediate vector $\mathbf{x}_t$ and then invert the process using the deterministic scheduler.

The Random Seeds baseline provides a naive reference point; by contrast, Diverse Diffusion and Particle Guidance are sampling-based methods explicitly designed to promote diversity, and serve as non-evolutionary benchmarks. Since Particle Guidance operate through classifier-free guidance, it is inapplicable to one-step models and are therefore only evaluated on Flux. ImageBreeder allows for comparison to diffusion-specific evolutionary operators, while Gaussian Noise and IsoLine represent standard evolutionary baselines. The Intermediate-Layer Gaussian baseline is critical for demonstrating that meaningful diversity results from spectral mutation specifically, rather than performing mutation at intermediate layers.

Hyperparameters For SDXL-Turbo, a computational budget of 2000 evaluations is allocated to all algorithms; for QD algorithms, this is processed as batches of 16 over 125 iterations. Flux experiments used a reduced computational budget of 400 total evaluations, processed as 100 iterations with a batch size of 4. Statistical significance is determined using a Wilcoxon signed-rank test with Holm-Bonferroni correction (Wilcoxon, 1992; Holm, 1979). Further hyperparameter details and an ablation study on these choices are provided in Appendices Section C.1 and Section D respectively.

4.3 Evaluation Metrics

Our primary metric is the Fréchet Inception Distance (FID) which is a widely-used metric that measures both the quality and diversity of solutions, compared to a set of real images (Heusel et al., 2017). This metric is not reported for algorithms that fail to discover a significant number of solutions. We provide further implementation details in Appendix Section C.6. Since the FID metric aggregates the Fitness and Diversity of solutions into one, we provide further insight via the following metrics:

- **Total Fitness:** The total fitness scores of all solutions in the final collection, reflecting both the quality of individual solutions and the algorithm's ability to populate the archive with solutions exceeding the fitness threshold.
- **Diversity (Mean Pairwise Distance):** The average Euclidean distance between all pairs of embeddings in the final collection, quantifying the semantic spread of the discovered portfolio.
- **Number of Solutions:** The total count of distinct solutions maintained in the final archive.

5 RESULTS

5.1 FRAMEWORK RESULTS

We first evaluate our framework using the FID score, as reported in Table 1. Our Spectral-Guided approach achieves the lowest FID scores across both SDXL-Turbo (81.90) and Flux (137.58) models. As FID measures the distance between the distribution of generated images and real-world data, these results confirm that our method produces high-quality, realistic outputs. Notably, baseline methods such as CVT-ME and CMA-MAE failed to discover a sufficient number of solutions across all runs for a valid FID metric to be calculated.

Table 1: FID Score.

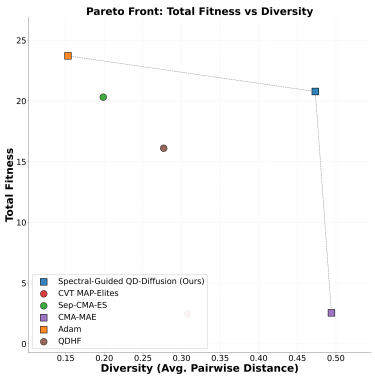
Algorithm	**SDXL-Turbo**	**Flux**
Spectral-Guided (Ours)	**81.90**	**137.58**
CVT MAP-Elites	n/a	n/a
Sep-CMA-ES	145.13	196.67
CMA-MAE	n/a	n/a
Adam	163.70	-
QDHF	122.20	-

Table 2: Number of Solutions.

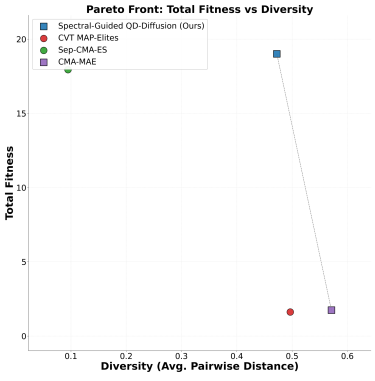
Algorithm	**SDXL-Turbo**	**Flux**
Spectral-Guided (Ours)	**62.7 ± 8.0**	**58.0 ± 17**
CVT MAP-Elites	6.9 ± 3.1	5.25 ± 3.4
Sep-CMA-ES	**61.4 ± 10.7**	**54.0 ± 23**
CMA-MAE	8.28 ± 2.3	5.92 ± 1.8
Adam	**64.0 ± 0.0**	-
QDHF	48.8 ± 7.8	-

To disentangle the contributions of diversity and quality, we examine the Total Fitness and Diversity metrics visualised via the Pareto Front in Figure 5. Compared to CVT-ME, our method exhibits clear Pareto domination on SDXL-Turbo, which attains lower Total Fitness (2.45 vs 20.8, $p < 2 \times 10^{-10}$) and a lower Diversity (0.31 vs 0.47, $p < 8 \times 10^{-7}$) than our method. On Flux, CVT-ME achieves a marginally higher diversity (0.497 vs 0.472) but the difference is not statistically significant ($p = 1.0$) and the total fitness is still substantially lower 19.5 vs 1.97, $p < 3.5 \times 10^{-2}$. This performance is attributable to its inability to densely populate the solution space, producing only 6.9 solutions per prompt on average and highlighting the limitations of grid-based QD methods in high-dimensional feature spaces.

Similarly, CMA-MAE also suffers from the grid-based approach, achieving a significantly lower Total Fitness on both models ($p < 4 \times 10^{-3}$). While it records higher diversity than our method, the difference is not statistically significant ($p = 0.47$ for SDXL-Turbo and $p = 0.389$ for Flux) and we hypothesise that this is largely a byproduct of poor image quality. As shown in the qualitative results in Appendix Section E.1, for SDXL-Turbo, many images are visually corrupted and for Flux, the model fails to maintain prompt adherence. For example, on the prompt "a glass of orange juice with an orange peel stuck on the rim", CMA-MAE generates images of dogs. This indicates that its recorded diversity stems from corrupted or irrelevant outputs rather than genuine semantic variation, which alongside its low solution count (e.g., 5.92 on Flux) explains its low Total Fitness scores.



(a) SDXL-Turbo.



(b) Flux.

Figure 5: Quality vs Diversity.

The presence of visual artefacts is also evident in the fitness-maximising algorithms, Adam and Sep-CMA-ES. A qualitative examination of images shows that both methods show signs of "reward hacking" by introducing inconsequential artefacts to artificially inflate fitness scores. In contrast, our approach discovers distinct interpretations such as varying viewing angles, background contexts, and compositions, while maintaining high prompt alignment. This is confirmed by the quantitative metrics: Adam (evaluated only on SDXL-Turbo) achieves the highest Total Fitness (23.71, $p < 3 \times 10^{-9}$) but suffers from a low Diversity score (0.15). Similarly, Sep-CMA-ES attains a Total Fitness comparable to our method on both SDXL-Turbo (20.3 vs 20.8, $p > 0.35$) and Flux (18.7 vs 19.5, $p > 0.6$). However, its diversity remains significantly lower on both (0.20 vs 0.47, $p < 4 \times 10^{-10}$, for SDXL-Turbo, 0.095 vs 0.472, $p < 6 \times 10^{-3}$ for Flux). These results highlight that pure fitness maximisation often stagnates in local optima, converging to a single mode of the design space regardless of the base model used.

Finally, compared to QDHF (evaluated only on SDXL-Turbo), our method achieves significantly higher fitness (16.1 vs 20.8, $p < 4 \times 10^{-7}$) and diversity (0.28 vs 0.47, $p < 4 \times 10^{-8}$). This suggests that projecting high-dimensional embeddings into lower-dimensional spaces may result in information loss and limit the algorithm's ability to retain diverse, high-performing solutions.

5.2 Variation Operator Results

We next isolate the contribution of the h-space Spectral Mutation operator. Table 3 confirms that our method achieves the lowest FID score compared to all other variation operators on both SDXL-Turbo and Flux, confirming its efficacy at balancing quality and diversity.

Table 3: FID Score.

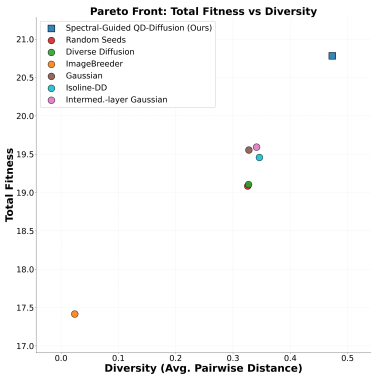
Algorithm	SDXL-Turbo	Flux
Spectral-Guided (Ours)	**81.90**	**137.58**
Random Seeds	100.12	180.03
Diverse Diffusion	104.59	185.94
Particle Guidance	-	186.74
ImageBreeder	207.75	247.03
Gaussian	93.34	144.14
Isoline-DD	91.52	146.65
Intermediate Gaussian	109.73	163.31

Table 4: Number of Solutions.

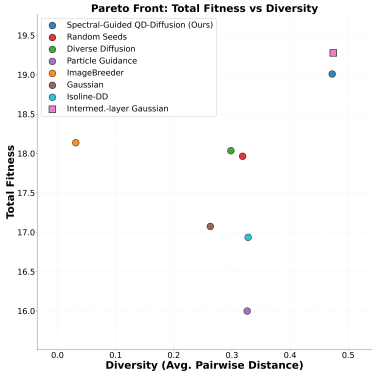
Algorithm	SDXL-Turbo	Flux
Spectral-Guided (Ours)	**62.7 ± 8.0**	**58.0 ± 17**
Random Seeds	**58.2 ± 15.0**	**54.0 ± 23**
Diverse Diffusion	**58.1 ± 15.0**	**54.0 ± 23**
Particle Guidance	-	**50.0 ± 25**
ImageBreeder	53.3 ± 20.2	**54.0 ± 23**
Gaussian	**58.7 ± 15.3**	**54.0 ± 23**
Isoline-DD	**58.7 ± 15.3**	**54.0 ± 23**
Intermediate Gaussian	**60.0 ± 13.5**	**59.0 ± 17**

Figure 6 shows that on both SDXL-Turbo and Flux, our method Pareto-dominates all operators across both models except the Intermediate-Gaussian on Flux. On Flux, the difference is statistically significant for both metrics for the Random Seeds and Imagebreeder baselines. Otherwise, for all other baselines on both models, the difference in fitness is not statistically significant for the fitness metric ($p > 0.05$), but it is for the diversity metric. This confirms that traditional variation operators are inefficient at traversing high-dimensional latent spaces compared to our spectral approach.

Finally, we compare our approach to the Intermediate-layer Gaussian baseline. Our method achieves significantly higher fitness on SDXL-Turbo (20.78 vs 19.59, $p < 5 \times 10^{-5}$) and diversity (0.473 vs 0.341, $p < 3 \times 10^{-10}$). On Flux, the Intermediate-layer Gaussian records a higher diversity score (0.498 vs 0.489) and comparable Total Fitness (19.46 vs 19.85), but the difference is not significant in either case ($p > 0.7$ for both). Moreover, a qualitative examination in Appendix Section E.2 reveals that for both models, the Intermediate-layer Gaussian produces heavily corrupted images. This indicates that its diversity arises from visual artefacts rather than genuine semantic variety, reinforcing the strength of the directional guidance provided by our h-layer spectral operator.



(a) SDXL-Turbo.



(b) Flux.

Figure 6: Quality vs Diversity.

6 Conclusion and Future Works

In this work, we introduced *Spectral-Guided Quality-Diversity Diffusion*, a framework designed to repurpose pre-trained diffusion models as searchable landscapes for evolutionary optimisation. By exploiting the spectral properties of the generative prior through our novel h-space Spectral Mutation operator, our method effectively navigates high-dimensional latent manifolds to uncover diverse, high-fidelity portfolios of solutions. Our empirical validation was conducted on the domain of text-to-image generation, where our method consistently outperforms 12 baselines across two architectures in terms of FID score, demonstrating a superior balance of image quality and diversity. While the framework was designed to be domain-agnostic, this cross-domain generalisability remains to be validated empirically. Future work will therefore look to extend this framework beyond images to scientific discovery, specifically applying spectral mutations to biological sequence design.

AI Assistance Statement

In this work, we used generative AI tools to aid and polish writing, for retrieval and discovery of related literature, to support research ideation and execution, and to draft sections of the paper text. We have not used generative AI tools to generate synthetic datasets or to prove mathematical claims; these are not applicable to this work.

We have reviewed all AI-assisted work. For example, when using AI to suggest related work, we read the corresponding literature to verify its relevance and manually added the references to our bibliography. Any drafts and edits of the text from AI was proofread for accuracy. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

R. Bahlous-Boldi, M. Faldor, L. Grillotti, H. Janmohamed, L. Coiffard, L. Spector, and A. Cully. *Dominated Novelty Search: Rethinking Local Competition in Quality-Diversity*, page 104–112. Association for Computing Machinery, New York, NY, USA, 2025. ISBN 9798400714658. URL https://doi.org/10.1145/3712256.3726310.

P. Bontrager, W. Lin, J. Togelius, and S. Risi. Deep interactive evolution. In *International Conference on Computational Intelligence in Music, Sound, Art and Design*, pages 267–282. Springer, 2018.

K. Chatzilygeroudis, A. Cully, V. Vassiliades, and J.-B. Mouret. Quality-diversity optimization: a novel branch of stochastic optimization. In *Black Box Optimization, Machine Learning, and No-Free Lunch Theorems*, pages 109–135. Springer, 2021.

M. J. Chong and D. Forsyth. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6070–6079, 2020.

L. Clare and J. Correia. Generating adversarial examples through latent space exploration of generative adversarial networks. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, pages 1760–1767, 2023.

G. Corso, Y. Xu, V. De Bortoli, R. Barzilay, and T. Jaakkola. Particle guidance: non-iid diverse sampling with diffusion models. *arXiv preprint arXiv:2310.13102*, 2023.

V. Costa, N. Lourenço, J. Correia, and P. Machado. Exploring generative adversarial networks for text-to-image generation with evolution strategies. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, pages 271–274, 2023.

A. Cully and Y. Demiris. Quality and diversity optimization: A unifying modular framework. *IEEE Transactions on Evolutionary Computation*, 22(2):245–259, 2018. ISSN 1941-0026. doi: 10.1109/TEVC.2017.2704781.

N. Deckers, J. Peters, and M. Potthast. Manipulating embeddings of stable diffusion prompts. *arXiv preprint arXiv:2308.12059*, 2023.

T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*, 2022.

L. Ding, J. Zhang, J. Clune, L. Spector, and J. Lehman. Quality diversity through human feedback: Towards open-ended diversity-driven optimization. *arXiv preprint arXiv:2310.12103*, 2023.

J. Dona, A. Flajolet, A. Marginean, A. Cully, and T. Pierrot. Quality-diversity for one-shot biological sequence design. ICML.

P. Fernandes, J. Correia, and P. Machado. Evolutionary latent space exploration of generative adversarial networks. In *International Conference on the Applications of Evolutionary Computation (Part of EvoStar)*, pages 595–609. Springer, 2020.

M. Fontaine and S. Nikolaidis. Differentiable quality diversity. *Advances in Neural Information Processing Systems*, 34:10040–10052, 2021.

M. Fontaine and S. Nikolaidis. Covariance matrix adaptation map-annealing. In *Proceedings of the genetic and evolutionary computation conference*, pages 456–465, 2023.

M. C. Fontaine, R. Liu, A. Khalifa, J. Modi, J. Togelius, A. K. Hoover, and S. Nikolaidis. Illuminating mario scenes in the latent space of a generative adversarial network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5922–5930, 2021.

S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

N. Hansen and A. Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Proceedings of IEEE international conference on evolutionary computation*, pages 312–317. IEEE, 1996.

A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

A. Inch, P. Chaiyapattanaporn, Y. Zhu, Y. Lu, T.-W. Ko, and D. Paglieri. Evolve to inspire: Novelty search for diverse image generation. *arXiv preprint arXiv:2511.00686*, 2025.

H. Janmohamed, M. Wolinska, S. Surana, T. Pierrot, A. Walsh, and A. Cully. Multi-objective quality-diversity for crystal structure prediction. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1273–1281, 2024.

T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

D. P. Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

G. Kwon and J. C. Ye. Improving diffusion-based image translation using asymmetric gradient guidance, 2023. URL `https://arxiv.org/abs/2306.04396`.

M. Kwon, J. Jeong, and Y. Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

B. F. Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll'ar, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023.

J.-B. Mouret and J. Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

D. P. Neto, J. Correia, and P. Machado. Evolutionary optimization trumps adam optimization on embedding space exploration. *arXiv preprint arXiv:2511.03913*, 2025.

A. M. Nguyen, J. Yosinski, and J. Clune. Innovation engines: Automated creativity and improved stochastic optimization via deep learning. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, pages 959–966, 2015.

Y.-H. Park, M. Kwon, J. Choi, J. Jo, and Y. Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36: 24129–24142, 2023.

J. K. Pugh, L. B. Soros, and K. O. Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40, 2016.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

M. Resnick, B. Myers, K. Nakakoji, B. Shneiderman, R. Pausch, T. Selker, and M. Eisenberg. Design principles for tools to support creative thinking. 2005.

B. Roziere, F. Teytaud, V. Hosu, H. Lin, J. Rapin, M. Zameshina, and O. Teytaud. Evolgan: Evolutionary generative adversarial networks. In *Proceedings of the asian conference on computer vision*, 2020.

S. Sadat, J. Buhmann, D. Bradley, O. Hilliges, and R. M. Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling, 2024. *URL https://arxiv. org/abs/2310.17347*, 6(7), 2024.

M. Salvenmoser and M. Affenzeller. Evolving the embedding space. In P. Machado, C. Johnson, and I. Santos, editors, *Artificial Intelligence in Music, Sound, Art and Design: 14th International Conference, EvoMUSART 2025*, volume 15611 of *Lecture Notes in Computer Science*, pages 402–416. Springer Nature, 2025. doi: 10.1007/978-3-031-90167-6_27.

A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024.

J. Simon. Creative exploration with reasoning llms, 2025. URL `https://www.joelsimon.net/lluminate`. Project page for "Lluminate," an LLM-based evolutionary algorithm.

D. Sobania, M. Briesch, and F. Rothlauf. *ImageBreeder: Guiding Diffusion Models with Evolutionary Computation*, page 463–471. Association for Computing Machinery, New York, NY, USA, 2025. ISBN 9798400714658. URL `https://doi.org/10.1145/3712256.3726439`.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

B. Tjanaka, M. C. Fontaine, D. H. Lee, Y. Zhang, N. R. Balam, N. Dennler, S. S. Garlanka, N. D. Klapsis, and S. Nikolaidis. Pyribs: A bare-bones python library for quality diversity optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '23, page 220–229, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701191. doi: 10.1145/3583131.3590374. URL `https://doi.org/10.1145/3583131.3590374`.

N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1921–1930, 2023.

V. Vassiliades and J.-B. Mouret. Discovering the elite hypervolume by leveraging interspecies correlation. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 149–156, 2018.

V. Vassiliades, K. Chatzilygeroudis, and J.-B. Mouret. Using centroidal voronoi tessellations to scale up the multidimensional archive of phenotypic elites algorithm. *IEEE Transactions on Evolutionary Computation*, 22(4):623–630, 2018. doi: 10.1109/TEVC.2017.2735550.

A. Vinhas, F. Assunção, J. Correia, A. Ekárt, and P. Machado. Fitness and novelty in evolutionary art. In *International Conference on Evolutionary and Biologically Inspired Music and Art*, pages 225–240. Springer, 2016.

J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

F. Wilcoxon. *Individual comparisons by ranking methods*. Springer, 1992.

M. Wolinska, A. Walsh, and A. Cully. Illuminating the property space in crystal structure prediction using quality-diversity algorithms. *Digital Discovery*, 3(8):1554–1563, 2024.

J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

M. Zameshina, O. Teytaud, and L. Najman. Diverse diffusion: Enhancing image diversity in text-to-image generation. *arXiv preprint arXiv:2310.12583*, 2023.

SUPPLEMENTARY MATERIALS

A PSEUDOCODE

A.1 MAP-ELITES PSEUDOCODE

Algorithm 1 MAP-Elites

Input: Search space $\mathcal{X}$, Feature function ϕ, Fitness function f, Archive $\mathcal{A}$ with empty cells corresponding to discretised features
Initialize archive $\mathcal{A}$ with random solutions sampled from $\mathcal{X}$
while not converged **do**
  Select parent x uniformly at random from $\mathcal{A}$
  Generate offspring $x' \leftarrow \text{Mutate}(x)$
  Evaluate fitness $f(x')$ and feature $\phi(x')$
  Identify cell c corresponding to $\phi(x')$
  if cell c is empty **or** $f(x') > f(\mathcal{A}[c])$ **then**
    $\mathcal{A}[c] \leftarrow x'$
  end if
end while
Return Archive $\mathcal{A}$

A.2 DNS Pseudocode

Algorithm 2 Dominated Novelty Search (DNS)

Input: Search space $\mathcal{X}$, Feature function ϕ, Fitness function f, Population size N, Neighbours k
Initialize population P with N random solutions sampled from $\mathcal{X}$
Evaluate fitness $f(x)$ and feature $\phi(x)$ for all $x \in P$
while not converged **do**
  $Q \leftarrow \emptyset$
  for $x \in P$ **do**
    $x' \leftarrow \text{Mutate}(x)$
    Evaluate $f(x')$ and $\phi(x')$
    $Q \leftarrow Q \cup \{x'\}$
  end for
  $R \leftarrow P \cup Q$
  for $x \in R$ **do**
    Find k-nearest neighbours of x in R with fitness $> f(x)$
    $\text{Score}(x) \leftarrow$ mean distance to these neighbours
  end for
  $P \leftarrow$ Select top N solutions from R based on Score
end while
Return Population P

B Related Works

B.1 Evolutionary Algorithms in Generative Design

Several works have demonstrated that Evolutionary Algorithms can outperform gradient-based methods like Adam (Kingma, 2014) when optimising the latent spaces of generative models such as GANs (Clare and Correia, 2023; Fernandes et al., 2020; Costa et al., 2023; Roziere et al., 2020; Nguyen et al., 2015). This observation extends to diffusion models, where EAs have been shown to locate higher-quality solutions than gradient descent (Neto et al., 2025; Salvenmoser and Affenzeller, 2025). However, the majority of these works focus on single-objective fitness maximisation rather than explicitly optimising for diversity via Quality-Diversity (QD) algorithms.

While some approaches do search for diverse imagery, they often operate under significant constraints. For instance, Vinhas et al. (2016) achieve diversity using Context Free Grammars, but their reliance on an unlimited archive classifies the method as novelty search rather than QD, and it does not operate within the high-dimensional embedding spaces typical of modern generative models. Similarly, Salvenmoser and Affenzeller (2025) employ an island-based Genetic Algorithm for open-ended search; however, this approach necessitates pre-defined artist styles per island, rendering it less generalisable.

Additionally, most of the existing works on using EAs for latent-space optimisation rely on traditional genetic operators, such as isotropic Gaussian noise. An exception is ImageBreeder (Sobania et al., 2025), which proposes specialized crossover and mutation operators for images. These include latent blending, channel-wise cuts, and mixed spatial cuts, all designed to recombine latent noise vectors from multiple parents. However, these operators remain largely agnostic to the geometry of the generative model. Alternatively, while some methods evolve text prompts directly (Inch et al., 2025; Simon, 2025), we avoid this to maintain a domain-agnostic framework applicable beyond language-conditioned tasks.

B.2 Quality-Diversity for Generative Design

Quality-Diversity algorithms have achieved notable success in illuminating the latent spaces of generative models. Early applications coupled QD with GANs to generate diverse content, such as video game levels (e.g., Mario scenes) (Fontaine et al., 2021) or face variations (Fontaine and Nikolaidis, 2023; 2021). A limitation of these works is their reliance on hand-defined, low-dimensional

features—for example, the "number of enemies" in a level or "hair length" in a face. Our approach, by contrast, utilises a high-dimensional feature embedding space, ensuring generalisability to domains like biological sequence design where such semantic features are not easily defined. Additionally, while methods like CMA-MAE (Fontaine and Nikolaidis, 2023) leverage gradients of the fitness or feature to accelerate search, our method exploits the spectral properties of the generative model itself to guide exploration.

B.3 Quality-Diversity for High-Dimensional Feature Spaces

Scaling QD to the high-dimensional outputs of modern generative models remains non-trivial. The standard solution for high-dimensional features is CVT-ME (Vassiliades et al., 2018), which uses Centroidal Voronoi Tessellations to discretise the feature space. However, this approach suffers from the curse of dimensionality, leaving many cells empty in high-dimensional grids. Dominated Novelty Search (DNS) (Bahlous-Boldi et al., 2025) addresses this by maintaining an unstructured archive. While effective in robotics tasks where offspring solutions remain in-distribution, we find that DNS requires modification for generative models to avoid retaining "hallucinated" out-of-distribution artifacts. Our method builds upon DNS by introducing a fitness thresholding mechanism to mitigate this issue.

Other approaches circumvent high-dimensional spaces by either a) using hand-defined metrics for diversity or b) incorporating human or LLM feedback to assess diversity (Simon, 2025; Bontrager et al., 2018). Notably, Quality-Diversity via Human Feedback (QDHF) (Ding et al., 2023) learns a projection layer (via contrastive loss from human/LLM feedback) to map high-dimensional embeddings into a manageable low-dimensional grid. While effective, we demonstrate that directly optimising within the high-dimensional embedding space reduces the information loss associated with projection, leading to superior archive quality and diversity.

B.4 Image Editing in Diffusion Models

To design a more effective mutation operator, we draw inspiration from diffusion-based image editing. Many works in this field use gradient-based optimization to maximize objectives like aesthetics (Deckers et al., 2023) or to alter an image towards a target text while preserving style (Kwon and Ye, 2023; Tumanyan et al., 2023). However, in contrast to our approach, these methods typically optimize for a specific "reward" or target, rather than exploring the space open-endedly.

Foundational analysis of the diffusion bottleneck, or h-space, has revealed that internal representations encode rich semantic geometry (Kwon et al., 2022; Park et al., 2023). Our mutation operator is heavily inspired by work analysing the Riemannian Geometry of the latent space of diffusion models Park et al. (2023). This work demonstrated that semantic edits can be generated by utilising a pullback metric to analyse the latent geometry of the latent-space. While this approach was developed for single-image *editing*, we repurpose these geometric insights (specifically the spectral analysis of the Jacobian) as a mutation operator for population-based search. Moreover, we extend this analysis beyond the latent noise vector to encompass all inputs of the diffusion model (including conditioning embeddings) considering them all as a genotype. Finally, we note the relevance of techniques such as Prompt-to-Prompt (Hertz et al., 2022) and Null-Text Inversion (Mokady et al., 2023) which established the importance of deterministic inversion mechanisms for preserving structure. These mechanisms are fundamental to our approach; without the ability to deterministically invert the diffusion process, modifications made at semantically rich intermediate layers could not be mapped back to the initial noise space to form a stable, evolvable genotype.

B.5 Image Diversity in Diffusion Models

Beyond evolutionary strategies, several inference-time techniques have been proposed to mitigate the "mode collapse" or lack of variety often observed in diffusion models. Diverse Diffusion (Zameshina et al., 2023) addresses this by optimising the dispersion of initial latent noise vectors, utilising a rejection-sampling approach to ensure that a batch of generated images maintains a minimum pairwise distance in the latent space. Similarly, Particle Guidance (Corso et al., 2023) replaces i.i.d. sampling with joint sampling of a batch of particles, augmenting the reverse diffusion process with the gradient of a pairwise-similarity potential (e.g. a kernel over image features) so that particles repel one another

and spread out across the distribution. While effective for batch-level variety, this approach is limited to the stochasticity of the prior and does not actively explore a semantic feature space. Finally CADS (Condition-Annealed Diffusion Sampler) (Sadat et al., 2024) is an inference-time, guidance-based diversity method that increases variety by annealing scheduled noise on the conditioning signal during sampling (Sadat et al., 2024). CADS was not applied to SDXL-Turbo, as its gradual annealing of the conditioning signal requires a multi-step denoising trajectory that the few-step Turbo schedule does not provide; it was likewise excluded from Flux, whose guidance distillation eliminates the explicit conditional–unconditional guidance interaction that CADS depends on, rendering the method inapplicable in its intended form.

Unlike these methods, which focus on immediate batch variety, our archive-based approach cumulates semantic diversity over multiple iterations, allowing for a more thorough exploration of rare or distal feature regions. Moreover, by operating on the genotype rather than the sampling trajectory, our method extends diversity-optimisation to one-step models such as SDXL-Turbo, where traditional guidance is inapplicable.

C Implementation Details

In this section, we detail the implementation specifics and hyperparameters used across all experiments. Our code is implemented using the Pyribs library (Tjanaka et al., 2023) and is available at *Code hidden for anonymity, will be released upon acceptance.*.

C.1 Global Hyperparameters

For SDXL-Turbo, all algorithms were allocated a total computational budget of 2000 evaluations. For Quality-Diversity (QD) algorithms, this corresponded to 125 iterations with a batch size of 16 offspring.

To fully explore the generative capabilities of SDXL-Turbo, we defined the genotype to include the latent noise vectors, the conditional prompt embeddings, and the pooled prompt embeddings. At each iteration, the batch of 16 offspring was divided among these components: 5 solutions mutated the latent vectors, 5 mutated the prompt embeddings, and 6 mutated the pooled prompt embeddings. The sole exception is the ImageBreeder baseline, which relies on operators designed specifically for image or latent space; for consistency in generalisation, we restricted ImageBreeder to mutate only the latent noise vector (see Section C.3).

For all archive-based algorithms (all algorithms except Adam and Sep-CMA-ES), the maximum archive capacity was set to 64 solutions. The fitness threshold for Dominated Novelty Search (DNS) archives, including the passive archives used for reporting metrics on Adam and Sep-CMA-ES, was set to 0.3.

Due to the substantial computational cost associated with the Flux model, we restricted this evaluation to a reduced computational budget of 400 total evaluations, processed as 100 iterations with a batch size of 4. The evaluation was performed on a subset of 12 prompts, consisting of one seed from each category of the Parti Prompts benchmark (prompts 1, 4, 7, etc. in Appendix Section F).

To mitigate runtime overhead, we applied 8-bit quantisation to the transformer component of the pipeline using the `BitsAndBytesConfig` from Hugging Face (Dettmers et al., 2022). Regarding the mutation operator, since Flux is a multi-step diffusion model, we randomly sampled the intermediate timestep t_* from the interval $[0.2, 0.8]$ for the Jacobian calculation and we restricted the genotype to the latent embedding solely. To accommodate the precision effects of quantisation, we adjusted the floating-point epsilon parameter for the Power Iteration SVD calculation to 1×10^{-1}. All other hyperparameters—including the number of SVD iterations, the number of principal directions, and the fitness threshold—remained identical to the main experiments. Crucially, we did not tune the mutation strength α for this architecture, testing the method's robustness to hyperparameter settings.

All experiments on both models were run on a high-performance computing cluster. Each experimental run was allocated a single node with 1 NVIDIA L40S GPU (48 GB VRAM), 16 CPU cores, and 48 GB of system memory.

C.2 Spectral-Guided Diffusion Implementation

Exact computation of the Jacobian is computationally intractable for high-dimensional diffusion models. We therefore approximate the principal singular vectors using the Power Iteration method described by Park et al. (2023). We estimate the first 3 principal directions using a maximum of 5 iterations and a floating-point epsilon of 1×10^{-3}.

During the mutation step, for each offspring, we randomly sample a direction $\mathbf{v}$ from these top 3 estimated singular vectors. The genotype is then updated according to the rule $\mathbf{x}' = \mathbf{x} + \alpha\mathbf{v}$, where the step size α is sampled uniformly from the range $[-10, 10]$. These parameters were consistent across all three genotype components and were used for any baseline employing the h-space Spectral Mutation Operator.

C.3 Baseline Implementations

CVT-ME As a grid-based algorithm, CVT-ME requires a defined discretisation of the feature space. We employed a Centroidal Voronoi Tessellation with 64 centroids to match the archive size of other methods. The bounds of the feature space were determined empirically to be $[-0.25, 0.25]$. Due to the high computational cost of Flux, we did not evaluate CVT-ME on this model. We instead prioritised CMA-MAE as the state-of-the-art grid-based baseline. Since both methods utilise the same CVT archive structure, CMA-MAE provides a more state-of-the-art benchmark for archival performance within our limited computational budget.

Sep-CMA-ES We tuned the initial step size σ for each genotype component independently across 5 random seeds, finding $\sigma = 0.1$ to be optimal for all components. We ran three independent Sep-CMA-ES processes for each genotype in parallel, corresponding to the batch sizes of the QD algorithms (5, 5, and 6). Following the Pyribs implementation (Tjanaka et al., 2023), the number of parents selected at each iteration was set to half the batch size.

CMA-MAE Consistent with our other grid-based baselines, for CMA-MAE, we used a CVT archive with bounds of $[-0.25, 0.25]$. We employed the Improvement Ranker (ranker="imp") to prioritise solutions that discovered new cells or improved existing ones. The initial step size was set to $\sigma_0 = 0.5$, and we utilised the basic restart rule to manage convergence. We used all mutated all three genotypes in parallel, corresponding to the batch sizes of the QD algorithms (5, 5, and 6). The archive learning rate was set to 0.01 with a fitness threshold of 0.0.

Adam We tuned the learning rate across 5 seeds and selected $\alpha = 0.01$. Unlike the population-based methods, Adam updates a single solution iteratively. To maintain a comparable computational budget, we performed 2000 gradient steps. Notably, Adam was not run on the Flux model, as the memory requirements for backpropagation on Flux exceeded available hardware limits, resulting in Out-Of-Memory (OOM) errors.

QDHF We followed the open-source implementation of Quality-Diversity via Human Feedback (QDHF). The algorithm learns a projection layer mapping high-dimensional CLIP embeddings to a low-dimensional latent space. We set this latent dimension to 2 and trained the projection layer at iterations corresponding to 1%, 10%, 25%, and 50% of the total budget. Training used a minibatch size of 32 and a buffer of 1000 preference data points. For the initial warmup phase (iteration 0), where no solutions exist, we generated 200 random images with a latent sigma of 0.01. These 200 image evaluations was given in addition to the main algorithmic 2000 evaluations budget. We excluded QDHF from the Flux experiments because the 200 "free" warmup images required for Flux would constitute half of the main method's total computational budget, making the comparison prohibitively expensive.

Diverse Diffusion We implemented the "Max" variant of Diverse Diffusion (Zameshina et al., 2023), adapted to operate as an emitter within our experimental framework. For each slot in the batch, the algorithm samples $n_{\text{max}} = 10$ candidate latent vectors from a standard normal distribution and selects the one that maximises the minimum L_2 distance to all previously selected vectors in the current batch. Following the original implementation, we used the latent embeddings as genotypes.

Particle Guidance We implemented Particle Guidance (PG) (Corso et al., 2023), generating each batch of particles jointly so that they repel one another throughout denoising. Particles were initialised from independent standard-normal noise and shared the parent's prompt and pooled embeddings. Following the original work, we induced repulsion with an RBF kernel in latent space, using the median heuristic for the bandwidth and scaling the repulsive gradient by a coefficient of 0.5 and the current noise level; repulsion was disabled once $\sigma_t < 0.5$ to preserve detail near the end of sampling. We used 28 denoising steps and an embedded guidance scale of 3.5. Particle Guidance was not applied to SDXL-Turbo, as its repulsion mechanism requires the samples to interact across multiple denoising steps, which the few-step Turbo schedule does not provide.

ImageBreeder ImageBreeder introduces nine mutation operators for diffusion models. We utilised the 'mixed-all' operator strategy, which randomly samples an operator at each iteration. To ensure domain agnosticism, we excluded the pixel-level operators ('pixel-blend' and 'pixel-blend-with-score') and utilised only the latent-based operators. Unlike other baselines initialized with a single seed, ImageBreeder relies on crossover operations which require population diversity to function effectively. Consequently, to avoid bias from a stagnant initial population, we seeded ImageBreeder with two distinct random images.

Gaussian Noise Mutation We tuned the mutation strength (σ) for each genotype component separately over 5 seeds. The optimal parameters were found to be $\sigma = 0.1$ for latent vectors, and $\sigma = 0.4$ for both prompt and pooled prompt embeddings.

IsoLine-DD Variation Operator For the IsoLine operator, we used the same isotropic σ_1 values as the Gaussian Noise baseline. We tuned the line mutation strength (σ_2) and found $\sigma_2 = 0.1$ to be optimal for all genotype components.

Intermediate Layer Gaussian To isolate the effect of spectral guidance, this baseline applies Gaussian noise at the intermediate h-layer rather than using singular vectors. Tuning over 5 seeds yielded an optimal $\sigma = 0.4$ for all three genotype components.

C.4 Baselines Evaluation

We evaluate QD baseline algorithms by computing the evaluation metrics established in Section 4 directly on the solutions in their final archives. For non-QD algorithms (i.e., CMA-ES, Adam, Random Seeds), we maintain a passive DNS archive with thresholding to calculate metrics comparably. This passive archive collects the solutions generated by the algorithm at each generation and is used exclusively for metrics reporting; it does not influence the main algorithm's search process. To ensure that performance differences arise from algorithmic dynamics every algorithm was seeded with the identical starting image for each prompt.

C.5 Parti Prompts Distribution

Figure 7 shows the Category and Challenge Distribution of prompts selected randomly from the Parti Prompts Dataset. We sampled 3 prompts from each of the categories leading to a total of 36 prompts (see Appendix Section F for the full list). The most common Challenge levels were "Basic" and "Fine-grained Detail".

C.6 FID Metric Implementation

To evaluate the visual quality and diversity of the generated archives, we calculate the Fréchet Inception Distance (FID) metric. This metric is computed using high-level feature representations extracted from the pre-trained Inception-v3 network (Szegedy et al., 2016), with the MS-COCO 2017 validation set serving as the reference distribution for real-world imagery (Lin et al., 2014).

C.6.1 FID Extrapolation

Since the FID metric is sensitive to sample size, we aggregate the images from all experimental runs and seeds for each baseline into a single combined archive. This accumulation is performed

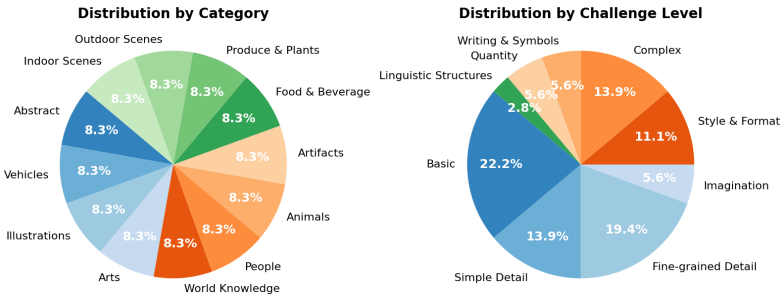


Figure 7: Pie charts representing the Category and Challenge Distribution of the sampled prompts.

specifically to maximize the available sample size, thereby reducing statistical variance and providing a more robust representation of each algorithm's output distribution. Despite this aggregation, the total archive sizes still vary between methods. We therefore adopt the FID_∞ extrapolation technique proposed by Chong and Forsyth (2020) to ensure an unbiased comparison.

Following the observation that the expected FID value is approximately linear with respect to the inverse of the number of samples N, we compute the raw FID at multiple subsampled intervals: $N \in \{200, 400, \ldots, 2000\}$. We then perform a linear least-squares regression of these values against $1/N$. The reported FID is the intercept of this regression line as $N \to \infty$, providing an estimate of the model's performance that is independent of the specific archive size. Algorithms that failed to produce the minimum required sample size for the regression analysis (e.g., $N < 200$) were excluded from the FID_∞ calculation, as a sufficient number of data points is mathematically necessary to produce a stable linear fit and avoid singular covariance matrices in the high-dimensional Inception feature space.

D Ablation Experiments

In this section, we include ablation studies on key hyperparameters of our method. Due to computational constraints, this study was performed only on SDXL-Turbo. The fitness threshold removal (Section D.1.1) and genotype ablation (Section D.2) were assessed using all 36 prompts, while the others used 12 prompts, one from each category.

D.1 Fitness Threshold Ablation

D.1.1 No Fitness Threshold

To validate the necessity of the fitness threshold mechanism within our method, we conduct an ablation study where this component is removed. Results for this comparison are organised in Table 5.

Table 5: Comparison between our full method and the fitness threshold ablation across diversity and fitness metrics.

Algorithm	Total Fitness	Avg. Pairwise Dist.	Num. Solutions
Spectral-Guided QD-Diffusion (Ours)	**20.8 ± 3.0**	0.473 ± 0.08	**62.7 ± 8.0**
Fitness Threshold Ablation	17.56 ± 1.8	**0.673 ± 0.07**	**64.0 ± 0.0**

As expected, the results demonstrate that while the ablation variant achieves a significantly higher diversity score ($p < 2 \times 10^{-10}$), it suffers from a Total Fitness score much lower than our full method ($p < 1 \times 10^{-2}$). This indicates an archive populated by solutions that are novel but ultimately invalid or semantically inconsistent.



(a) Fitness Threshold Ablation



(b) Spectral-Guided QD-Diffusion (Ours)

Figure 8: Qualitative comparison for the prompt "a knight holding a long sword". The ablation (a) exhibits high diversity but often fails to meet semantic constraints, whereas our method (b) maintains high fitness and validity.

This disparity is further clarified through qualitative analysis. Figure 8 displays the top-16 solutions for the prompt "a knight holding a long sword" generated by the ablation and our full method. While some of the images produced by the ablation are novel and interesting, the archive contains many low-quality examples. For instance, in one image the "knight" is rendered as an old man and his "sword" is a walking stick. Similar behaviour is observed for more complex prompts, such as "a drop-top sports car coming around a bend", where the ablation baseline tends to produce stretched car geometries, toy cars, or empty roads. While these results highlight the raw creativity of the h-space Spectral Mutation operator, they also underscore the critical necessity of our fitness gating mechanism to maintain semantic integrity.

D.1.2 Fitness Threshold Robustness

Beyond the binary removal of the threshold, we evaluate the robustness of our method to specific threshold choices by testing a range of values. The results are summarised in Table 6.

Table 6: Robustness analysis of the fitness threshold. Performance varies smoothly, with 0.3 serving as a stable default.

Threshold Value	Total Fitness	Avg. Pairwise Dist.	Num. Solutions
0.26	**20.090 ± 0.965**	**0.576 ± 0.069**	**64.000 ± 0.000**
0.28	**20.673 ± 1.100**	0.542 ± 0.068	**64.000 ± 0.000**
0.30 (Default)	**21.104 ± 1.143**	0.500 ± 0.072	**64.000 ± 0.000**
0.32	**17.647 ± 8.035**	0.344 ± 0.146	**52.000 ± 21.709**
0.34	**12.454 ± 9.789**	0.256 ± 0.137	**36.083 ± 24.645**

The system demonstrates significant robustness, with performance metrics varying smoothly across the tested range. Lowering the threshold below 0.3 allows for a slight increase in diversity by permitting more marginal solutions, while raising the threshold above 0.3 results in a monotonic decline in archive coverage and Total Fitness. This suggests that while the mechanism is robust, the default threshold effectively balances semantic integrity with diversity.

D.2 Genotype Ablation

In this section, we analyse the impact of different genotype components on the algorithm's performance. Our framework, as proposed in Section 3.1, is designed to be general, allowing any input to the generative model to serve as a mutable genotype. For generality, our main experiments employed a mixed strategy, mutating the latent noise, prompt embeddings, and pooled prompt embeddings in equal proportions. To verify the validity of this choice, we conducted an ablation study isolating each component. All other hyperparameters were identical to the main experiments (Section C.1).

Table 7: Performance comparison across algorithms. Bold values indicate the best performer or algorithms not significantly worse than the best ($p \geq 0.05$).

Algorithm	Total Fitness	Diversity	No. of Solutions
Mixed Genotype	**20.782 ± 3.013**	**0.473 ± 0.083**	**62.667 ± 8.000**
Latent Noise	**20.393 ± 4.238**	0.351 ± 0.093	**61.333 ± 11.151**
Prompt Embedding	19.988 ± 4.005	**0.480 ± 0.096**	**61.333 ± 11.151**
Pooled Embedding	18.792 ± 6.042	0.357 ± 0.150	**57.056 ± 16.805**

Table 7 presents the results averaged across all 36 prompts. We observe no statistically significant difference in the Number of Solutions across any of the methods, indicating all genotype strategies are capable of populating the archive. In terms of Total Fitness, the Mixed Genotype and Latent Noise approaches perform best, significantly outperforming the Pooled Prompt Embedding. However, regarding Diversity, both our Mixed Genotype strategy and the Prompt Embedding ablation achieve the highest scores, significantly surpassing the Latent Noise and Pooled Prompt Embedding approaches.

These results suggest a trade-off: mutating the Latent Noise is highly effective for maintaining fitness but yields lower diversity, while mutating the Prompt Embedding drives significant semantic variation (Diversity) at a slight cost to fitness. Our Mixed Genotype approach successfully leverages the strengths of both, achieving top-tier performance in both Total Fitness and Diversity. Consequently, we retain the mixed strategy for our final experiments to ensure robustness and generalisability across different generative tasks.

D.3 Mutation Strength (α) Ablation

We further investigate the sensitivity of our method to the mutation strength parameter, α, which dictates the magnitude of the h-space perturbation. Table 8 compares our default value against halving and doubling the magnitude.

Table 8: Ablation of mutation strength α. The method remains stable across varying perturbation magnitudes.

Algorithm	Total Fitness	Avg. Pairwise Dist.	Num. Solutions
$\alpha = 5$	**19.668 ± 5.020**	0.451 ± 0.070	**60.000 ± 13.856**
$\alpha = 10$ (Default)	**21.104 ± 1.143**	0.500 ± 0.072	**64.000 ± 0.000**
$\alpha = 20$	**21.087 ± 1.015**	**0.562 ± 0.059**	**64.000 ± 0.000**

The results indicate that doubling or halving the perturbation magnitude has a negligible effect on overall performance. All configurations achieved comparable fitness and diversity, suggesting that the Spectral Mutation operator is not hyper-sensitive to the exact scale of α.

D.4 Spectral Direction (k) Iteration Ablation

To determine the number of spectral directions required to effectively guide the mutation, we ablate the value of k in the SVD-based operator. Results are organised in Table 9.

Table 9: Impact of the number of spectral directions k on algorithm performance.

Algorithm	Total Fitness	Avg. Pairwise Dist.	Num. Solutions
$k = 1$	**18.617 ± 6.630**	**0.453 ± 0.083**	**56.615 ± 18.026**
$k = 3$ (Default)	**21.104 ± 1.090**	**0.495 ± 0.071**	**64.000 ± 0.000**
$k = 5$	**21.091 ± 1.104**	**0.480 ± 0.075**	**64.000 ± 0.000**
$k = 10$	**21.191 ± 1.158**	**0.465 ± 0.068**	**64.000 ± 0.000**

Performance effectively plateaus at $k = 3$, with diminishing returns observed for higher values. Only the reduction to $k = 1$ leads to a meaningful drop in archive stability and fitness. This suggests that a small number of spectral directions is sufficient to capture the semantically active subspace required for effective Quality-Diversity exploration.

D.5 Targeted Hidden Layer Ablation

Finally, we explore the impact of targeting different architectural blocks within the U-Net for spectral mutation. Table 10 compares our default choice (`mid_blocks`) against other potential targets.

Table 10: Comparison of spectral mutation performance when applied to different U-Net layers.

Target Layer	Total Fitness	Avg. Pairwise Dist.	Num. Solutions
`mid_blocks` (Default)	**21.104 ± 1.143**	**0.500 ± 0.072**	**64.000 ± 0.000**
`down_blocks.2`	**21.380 ± 1.170**	**0.499 ± 0.071**	**64.000 ± 0.000**
`up_blocks.1`	**20.021 ± 5.551**	**0.465 ± 0.093**	**59.636 ± 14.473**
`up_blocks.2`	**20.098 ± 5.596**	**0.467 ± 0.109**	**59.636 ± 14.473**

The choice of layer has a minimal overall impact on the results. However, we observe that targeting decoder-side layers (`up_blocks`) results in slightly lower stability and fitness compared to bottleneck or encoder-side layers, justifying our choice of the central blocks.

E Sample Image Results

E.1 Framework Experiment Images

Figure 9 and Figure 10 shows the images generated by our framework baselines and the corresponding seed image for sample prompts, for SDXL-Turbo and Flux respectively.

E.2 Variation Operator Experiment Images

In Figure 11 and Figure 12 we include the results images generated by our variation operator baselines and the seed image for sample prompts, for SDXL-Turbo and Flux respectively.

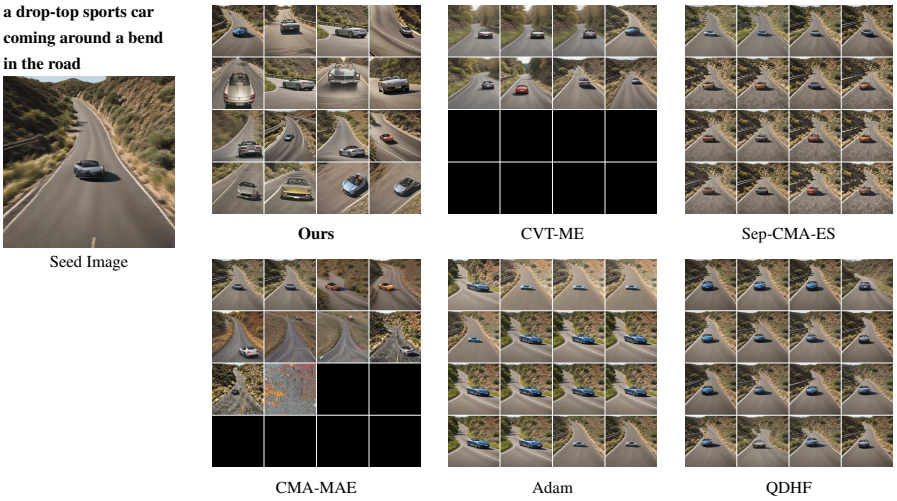


Figure 9: SDXL-Turbo Top-16 images with highest fitness for sample prompts.



Figure 10: Flux Top-16 images with highest fitness for sample prompts.



Figure 11: SDXL-Turbo Top-16 images with highest fitness for sample prompts for Variation Operator Experiments.

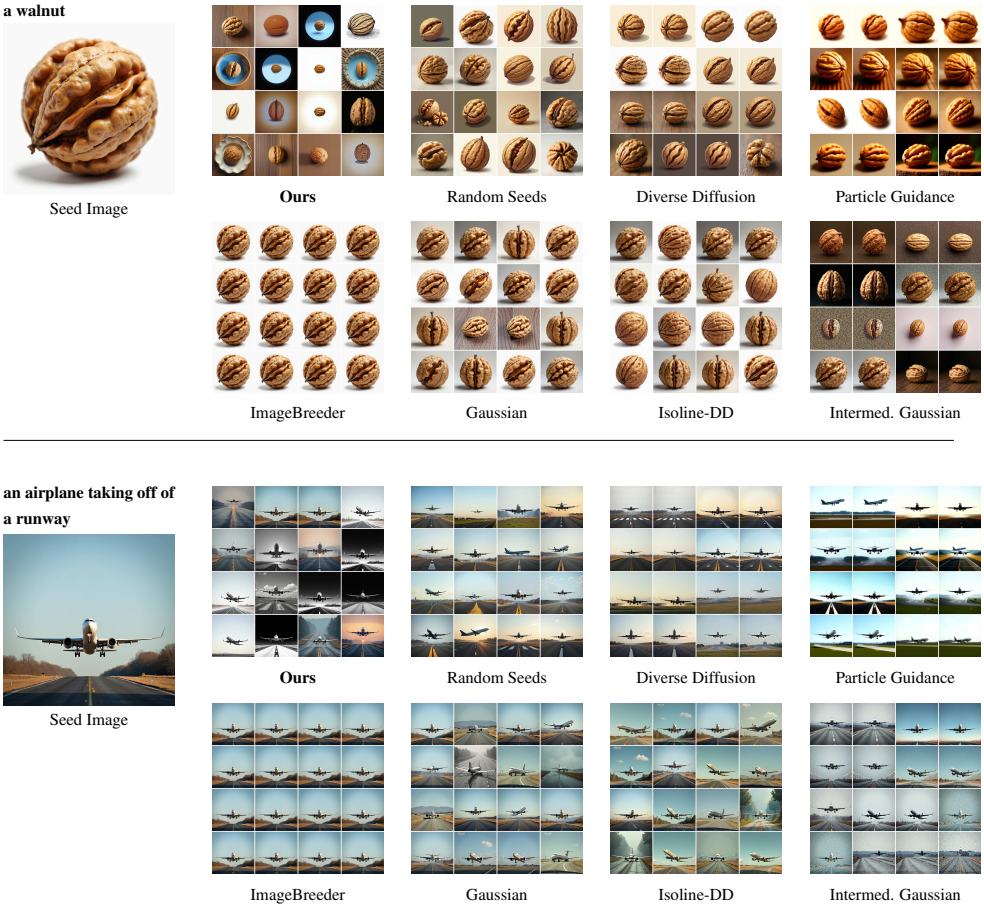


Figure 12: Flux top-16 images with highest fitness for sample prompts for Variation Operator Experiments.

F Full List of Prompts

Table 11: List of prompts with categories and challenge levels.

ID	Category	Challenge Level	Prompt
1	Abstract	Basic	happiness
2	Abstract	Basic	metal
3	Abstract	Basic	element
4	Vehicles	Simple Detail	an airplane taking off of a runway
5	Vehicles	Fine-grained Detail	an antique car by a beach
6	Vehicles	Fine-grained Detail	a drop-top sports car coming around a bend in the road
7	Illustrations	Imagination	a cube made of porcupine
8	Illustrations	Basic	a musical note
9	Illustrations	Style & Format	a sketch of a skyscraper
10	Arts	Complex	an oil surrealist painting of a dreamworld on a seashore where clocks and watches appear to be inexplicably limp and melting in the desolate landscape. a table on the left, with a golden watch swarmed by ants. a strange fleshy creature in the center of the painting
11	Arts	Complex	A raccoon wearing formal clothes, wearing a tophat and holding a cane. The raccoon is holding a garbage bag. Oil painting in the style of Vincent Van Gogh.
12	Arts	Style & Format	an abstract painting of a tree and a building
13	World Knowledge	Basic	The Statue of Liberty
14	World Knowledge	Basic	the grand canyon
15	World Knowledge	Complex	A portrait of a metal statue of a pharaoh wearing steampunk glasses and a leather jacket over a white t-shirt that has a drawing of a space shuttle on it.
16	People	Fine-grained Detail	a knight holding a long sword
17	People	Fine-grained Detail	an elder politician giving a campaign speech
18	People	Simple Detail	children on a couch
19	Animals	Writing & Symbols	Portrait of a tiger wearing a train conductor's hat and holding a skateboard that has a yin-yang symbol on it. woodcut
20	Animals	Basic	a cat
21	Animals	Style & Format	Portrait of a gecko wearing a train conductor's hat and holding a flag that has a yin-yang symbol on it. Oil on canvas.
22	Artifacts	Fine-grained Detail	a yellow diamond-shaped sign with a deer silhouette
23	Artifacts	Writing & Symbols	a black t-shirt with the peace sign on it
24	Artifacts	Simple Detail	a yield sign
25	Food & Beverage	Fine-grained Detail	a glass of orange juice with an orange peel stuck on the rim
26	Food & Beverage	Simple Detail	a roast turkey on the table
27	Food & Beverage	Imagination	A castle made of tortilla chips, in a river made of salsa. There are tiny burritos walking around the castle
28	Produce & Plants	Basic	a walnut
29	Produce & Plants	Style & Format	A photo of a palm tree made of water.
30	Produce & Plants	Complex	a tree reflected in the hood of a blue car
31	Outdoor Scenes	Simple Detail	a street with several cars on it
32	Outdoor Scenes	Quantity	two chemtrails forming an X in blue sky
33	Outdoor Scenes	Linguistic Structures	a house with no windows
34	Indoor Scenes	Complex	An empty fireplace with a television above it. The TV shows a lion hugging a giraffe.
35	Indoor Scenes	Fine-grained Detail	a wood cabin with a fire pit in front of it
36	Indoor Scenes	Quantity	a bunch of laptops piled on a sofa

G Wall-clock Time Experiments

In this section, we provide a detailed analysis of the computational cost of the h-space Spectral Mutation operator. This analysis was performed on 12 seeds, with one seed selected from each prompt category. We exclude CVT-ME and QDHF from this comparison as both utilise our Spectral Mutation operator as their primary variation mechanism.

G.1 Per-Image Generation Time

The median time per image generation for each method is provided in Table 12.While most evolutionary operators rely on simple additive noise, our approach involves a Jacobian-based spectral estimation which adds meaningful overhead relative to the base model's forward pass.

On SDXL-Turbo, our Spectral operator is the most expensive mutation other than Adam, which requires a full backward pass through the model. On Flux.1, our per-image cost ($14{,}564$s) is of the same order as the guidance-based diversity method, Particle Guidance ($18{,}885$s), and is roughly $2.5\times$

Table 12: Computational Overhead: Median Time per Image (Seconds)

Algorithm	SDXL-Turbo (s)	Flux (s)
Spectral-Guided (Ours)	0.78	14564.48
Sep-CMA-ES	0.002	7.22
CMA-MAE	0.001	6.96
Adam	0.90	-
Random Seeds	0.0003	1.37
Diverse Diffusion	0.002	16.39
Particle Guidance	-	18885.01
ImageBreeder	0.002	7.85
Gaussian	0.002	6.36
Isoline-DD	0.002	6.84
Intermed.-layer Gaussian	0.003	5864.6

that of the Intermediate-layer Gaussian operator ($5{,}865$s). It remains several orders of magnitude greater than the simple additive-noise operators, such as Random Seeds, Gaussian, and Isoline-DD, whose mutations cost only a few seconds per image.

We acknowledge that this overhead is a key limitation of our work: the Jacobian-based spectral estimation is substantially more costly than unstructured noise-based mutation. We note, however, that this expense is characteristic of methods that actively steer the generative trajectory rather than perturbing latents with unstructured noise. The baseline whose cost is comparable to ours, Particle Guidance, are precisely those designed to promote diversity through guidance, which suggests that the overhead largely reflects the general cost of structured, semantically guided exploration rather than an inefficiency specific to our operator. Whether this cost is warranted ultimately depends on the quality and diversity of the resulting solutions, which we examine under a fixed evaluation budget in Section 5.1 and Section 5.2. Reducing this overhead, for example through cheaper approximations to the spectral decomposition, is a promising direction for future work.

Finally, we note that the significance of this overhead is context-dependent. The per-image cost is most salient in settings where image generation is the principal bottleneck and candidates are cheap to evaluate. In many scientific design applications, however, the cost of generation is small relative to the cost of downstream validation: in domains such as biological or materials design, for instance, evaluating a candidate may involve wet-lab synthesis, assays, or physical testing that can dominate the overall time and resource budget. In such regimes, the additional generation cost of our operator is likely to be a minor contribution to the total, and the benefit of surfacing a more diverse set of candidates, thereby reducing the number of expensive validation cycles needed to find a viable solution, may well outweigh it. We do not evaluate such applications here, and the trade-off will depend on the specific costs involved, but we highlight this as a setting in which the computational profile of our method may be more favourable.

H Impact Statement

This paper presents work whose goal is to advance the field of machine learning, specifically at the intersection of Evolutionary Computation and Generative AI. Our proposed framework, Spectral-Guided Quality-Diversity, enables the systematic exploration of the latent spaces of large-scale generative models. This capability has significant potential for positive societal impact, particularly in scientific discovery, where it can accelerate the design of novel proteins, materials, or drugs by generating diverse portfolios of candidates that might otherwise be missed by mode-seeking methods. Furthermore, by explicitly optimising for diversity, our approach contributes to bias mitigation, helping to uncover regions of the data distribution that are often suppressed by standard sampling, potentially surfacing underrepresented groups or concepts and counteracting mode-collapse-induced bias.

However, we acknowledge that methods designed to uncover "hidden" diversity inherently carry risks. By navigating away from the high-likelihood modes of the training distribution, such algorithms could potentially circumvent safety alignment measures designed to suppress harmful, offensive, or copyrighted content. We encourage researchers and practitioners employing this framework

to integrate robust post-generation safety filtering, particularly when deploying these methods in open-ended or user-facing applications.